THE PRIMARY STRUCTURES OF NON-HISTONE CHROMOSOMAL PROTEINS HMG 1 AND 2

John M. WALKER+, Keith GOODERHAM*, Jeremy R. B. HASTINGS, Elaine MAYES and Ernest W. JOHNS

Chester Beatty Research Institute, Institute of Cancer Research, Royal Cancer Hospital, Fulham Road, London, SW3 6JB, England

Received 24 October 1980

1. Introduction

It is now generally accepted that in all eukaryotes the DNA and the histone proteins are complexed together in a repeating unit called the nucleosome (review [1]). The nucleosome consists of ~200 basepairs of DNA complexed with an octamer of histones H2A, H2B, H3 and H4, and interacts with 140–145 basepairs of DNA to form the core particle of the nucleosome. The fifth histone, H1, is associated with a variable length of spacer or linker DNA which separates the repeating units. The amino acid sequences of all 5 histones have now been known for a number of years. However, until recently very little sequence information has been available for any of the non-histone chromosomal proteins involved in the nucleosome structure. We have been studying a particular group of non-histone chromosomal proteins called the HMG proteins (review [2]). The presence of HMG proteins in a variety of organisms and tissues, including avian erythrocytes [3,4], trout testis and trout liver [5,6], wheat and yeast [7] and insects [8] implies a widespread occurrence in eukaryotic nuclei. There are 4 main HMG proteins in thymus, HMG 1, 2, 14 and 17. All 4 of these proteins have been isolated in a pure form from both pig and calf thymus [9–11], and have all been shown to be present in isolated nucleosomes [12]. The primary structures of both HMG 14 and 17 have been determined [13,14]. Because of the quantities of the HMG proteins present in the nucleus (10^5–10^6 molecules of each protein), we consider that the HMG proteins are structural proteins, possibly involved in the higher order structure of the chromatin, and not involved in specific gene control. There is evidence, however, that HMG 14 and 17 may be involved in the maintenance of the structure of active genes [15,16].

One of the noteworthy features of HMG proteins 1 and 2 is that >50% of their amino acid residues are charged. Like the histones, 25% of the residues in both proteins are basic. However, unlike the histones, both proteins also contain 30% acidic amino acids [2].

In [17,18] we have published sequence data for the cyanogen bromide peptides from both HMG 1 and 2. Here, we describe the production of peptides from HMG 1 and 2 by peptic cleavage of the native protein and by tryptic cleavage of the succinylated protein. Sequence data for these peptides, together with the data in [17,18], allow the description of the primary structures of both HMG 1 and 2.

2. Experimental

2.1. *Isolation of proteins*

Proteins HMG 1 and 2 were prepared as in [2].

2.2. *Pepsin cleavage of HMG 1 and 2*

Pepsin (1:2000, w/w) was added to protein (5 mg/ml in 5% acetic acid) and digestion allowed to continue for 1 h at 4°C. At the end of this time the reaction was stopped by the addition of pepstatin (10 × wt pepsin used, added in a small volume of DMSO). The digest mixture was then rotary evaporated to dryness and dissolved in 20 ml sodium ace-

+ Present address: School of Biological Sciences, Hatfield Polytechnic, Hatfield, Herts, England

* Present Address: MRC Clinical and Population Cytogenetics Unit, Western General Hospital, Crewe Road, Edinburgh, Scotland

tate buffer (0.01 M (pH 3.8), containing 0.5 ml β-mercaptoethanol/l). This solution was run onto a column of CM 52 cellulose (2.4 X 40 cm) equilibrated in the same buffer, at a flow rate of 45 ml/h. Peptides were eluted by the passage of a further 50 ml buffer followed by a salt gradient (2 X 700 ml, 0.1–0.6 M NaCl) in the same buffer. Eluted peptides were detected by their A_{230}. Pooled samples were rotary evaporated to a small volume (~5 ml) then desalted by passage through a column of Sephadex G-75 equilibrated in 0.01 N HCl. Eluted peaks were rotary evaporated to dryness.

2.3. *Succinylation and tryptic cleavage of HMG 1 and 2*

Protein was dissolved in water (10 mg/ml) and the pH adjusted to 7.0 with solid sodium carbonate. Solid succinic anhydride (1.0 g/100 mg protein) was added in small aliquots and the pH maintained at 7–7.5 by the addition of solid sodium carbonate. When all the succinic anhydride had been added the pH was raised to 8 by the addition of sufficient sodium carbonate and the solution left for 2 h. The succinylated protein was recovered by dropping the pH of the solution to 2 with hydrochloric acid. Precipitated protein was collected by centrifugation, washed once with 0.1 N HCl, then 3 times with acetone and dried under vacuum.

Succinylated protein was dissolved in ammonium bicarbonate solution (10 mg/ml, 0.2 M), trypsin (1:100, w/w) added, and the sample incubated at 37°C for 4 h, at which time the digest was stopped by the addition of soya bean trypsin inhibitor (equal w/w of trypsin used). The sample was rotary evaporated to dryness and dissolved in 20 ml 0.01 M ammonia/0.1 M NaCl, then run onto a column of QAE-A50 Sephadex (2.4 X 25 cm) equilibrated in the same buffer, at a flow rate of 30 ml/h. Peptides were eluted by the passage of 50 ml buffer followed by a salt gradient from 0.1–1.0 M NaCl (2 X 500 ml) in the same buffer. Eluted peptides were identified and desalted as described for the pepsin digest with the exception that the desalting step was carried out in 0.01 M ammonia.

2.4. *Peptide sequence determinations*

Automated Edman degradations of large peptides (>50 residues) were carried out on a Beckman 890C protein sequencer using a 0.1 M quadrol buffer programme with a double cleavage step on each cycle, essentially as in [19]. The same programme was used for small peptides (<50 residues) but with the addition that polybrene (5 mg) was used as a carrier and taken through 3 cycles of the Edman degradation together with 100 nmol glycyl glycine prior to each sequenator run [19]. Because of their highly charged nature it was not found necessary to use polybrene with any of the succinylated peptides.

PTH derivatives of released amino acids were determined both directly by high pressure liquid chromatography (HPLC) and indirectly by back-hydrolysis to the free amino acid. HPLC was carried out on a DuPont 830 Liquid Chromatogram using a Partisil PX5 ODS column (Whatman). PTH amino acids were eluted with a linear gradient of acetonitrile from 15–48% in 0.01 M sodium acetate buffer (pH 4.5) over 7 min, then holding at 48% acetonitrile for a further 4 min. Eluted PTH amino acids were identified by their A_{269}. Back-hydrolysis of PTH amino acids was carried out in 65% hydriodic acid at 110°C for 24 h. Liberated amino acids were identified on a Rank-Hilger Chromaspek amino acid analyser.

3. Results

3.1. *The primary structure of HMG 1*

The primary structure of HMG 1 is shown in fig.1 together with details of the major peptide sequences used to derive this sequence.

Ion-exchange chromatography of the products from tryptic cleavage of succinylated HMG 1 gave 5 major peaks but also a background of a number of minor peaks. It was obvious from the results obtained from this digest that tryptic cleavage did not occur exclusively at arginine residues. However, despite this, 3 major succinylated peptides were isolated which were of use for sequence studies. Peptide A was a 22 residue peptide running from residues 56–77. The total sequence of this peptide allowed the overlap of 2 CNBr peptides (peptides CB5-1 and CB5-2, residues 60–70 and 71–82) [17] and also extended the sequence at the N-terminus of one of these peptides (at residues 55–59). Peptide B provided the overlap between these 2 linked peptides and a further CNBr peptide (peptide CB3) at residues 81–83 (CB3 commences at residue 83) [17]. Peptide C was produced by cleavage at arginine at residue 169. Sequenator analysis of this peptide allowed 32 residues to be determined. This peptide starts towards the end of the N-terminal sequence of CNBr peptide

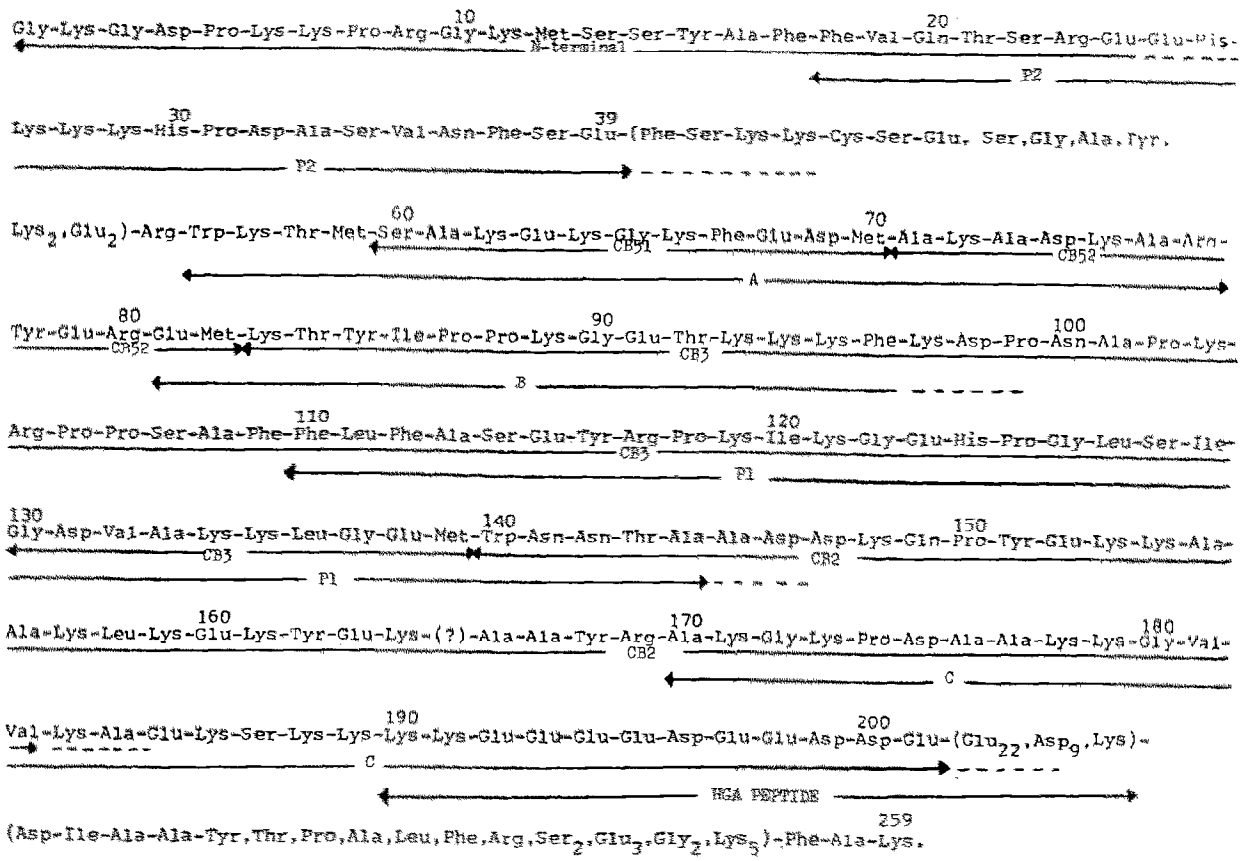

Fig.1. The primary structure of calf thymus HMG 1.

CB2 [17] and continues the CB2 sequence into a highly acidic sequence of aspartic and glutamic acid residues, termed the HGA region of the molecule (20). Because of the repetitive nature of this HGA sequence and the known difficulty of totally washing out thiazoline derivatives of glutamic acid from the reaction cup at a given step, hence producing overlaps, it has not proved possible to determine an unambiguous sequence for the HGA peptide beyond residue 201.

Peptic cleavage of HMG 1 gave 2 major peaks on ion-exchange chromatography peptide P1 eluting at ~0.3 M NaCl and peptide P2 eluting at 0.5 M NaCl. Analysis of peptide P2 showed that this peptide was produced essentially by cleavage at the Phe–Phe bond at residues 17 and 18, but also by a lesser cleavage at the Phe–Val bond at residues 18 and 19.

Sequenator analysis of P2 only produced 22 residues in sequence. The fact that only a short length of sequence was obtained was due to three factors:

(1) The 'staggered' N-terminus of the peptide giving overlap at each step;
(2) A reduced yield of PTH amino acid at and beyond residue 20 due, presumably, to partial cyclisation of glutamine at this step;
(3) The presence of 2 histidine residues at positions 26 and 30 which both showed the phenomenon of a 'histidine jump', where part of the sequence effectively 'jumps' one residue during sequenator analysis. This had the effect of exaggerating the stagger already present in the sequence.

This sequence did, however, extend the N-terminal sequence in [21] to residue 39. Amino acid analysis of P2 suggested that the peptide extended to residue 109 (peptide P1 commences at residue 110, see below). Tryptic, thermolytic and V8-protease cleavage of peptide P2 produced numerous peptides (not shown) which confirmed that P2 extended to residue 109 and provided peptides that covered the total sequence data from residue 55–110. Comparison of

our final sequence data with the amino acid analysis of P2 suggested a region of undetermined sequence between residues 39 and 55. V8-protease digestion of P2 had given one unplaced peptide;
Phe–Ser–Lys–Lys–Cys–Ser–Glu
which has been placed in this region. The same digest also gave the peptide:
Arg–Trp–Lys–Thr–Met–Ser–Ala–Lys–Glu (residues 55–63)
which suggests residue 54 in Glu. However, despite exhaustive studies of tryptic, V8-protease and thermolytic peptides from P2, peptides necessary to account for the 8 residues still remaining to be placed in sequence were not found. However, since both HMG 1 and 2 have complex isoelectric focusing patterns, and since the reason for this complexity has yet to be determined, it is possible that our inability to determine the sequence in this region is due to the fact that HMG 1 is microheterogeneous in this region, in much the same way at H1 has short microheterogeneous regions [22,23].

Analysis of peptide P1 showed peptic cleavage to nave occurred essentially at the Phe–Phe bond at residues 109–110, but minor cleavages also occurred at the Phe–Leu and Leu–Phe bonds between residues 110–112. Sequenator analysis of 35 residues of this peptide allowed the overlap of 2 major CNBr peptides CB3 and CB2 [17]. Amino acid analysis showed peptide P1 to extend to the C-terminus of the molecule. Tryptic and V8-protease digestion of peptide P1 provided peptides which covered the majority of the sequence from residues 110 up to and including the HGA sequence which starts at residue 192. Only one unplaced peptide, a tryptic peptide with the partial sequence Asp–Ile–Ala–Tyr was isolated from P1 and this has been placed after the HGA peptide. The C-terminal of the molecule had been determined as –Phe–Ala–Lys in [24]. Comparison of our sequence data with the amino acid analysis of P1 therefore suggests a further 18 residues remain to be placed in sequence after the HGA peptide. Again, the possibility that microheterogeneity exists in this

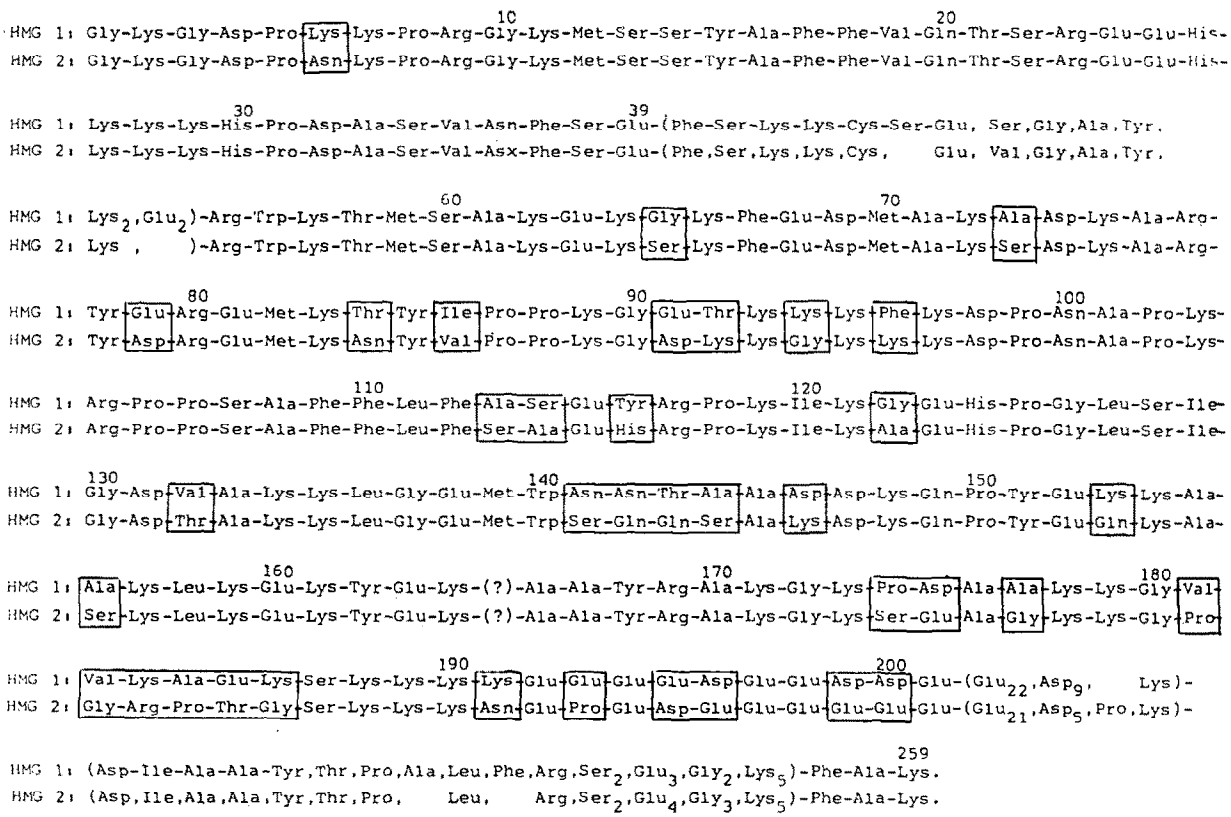

Fig.2. The primary structure of calf thymus HMG 1 and HMG 2.

region should be considered. A smaller peak, P0, which eluted just before peptide P1, was shown to be the N-terminal residues 1–17. Amino acid analysis of peptide P1 showed that the second cystein residue known to be in the molecule was present in this peptide. Since it was also known that peptide CB2 contains a cysteine residue this places the second cysteine residue between residue 140 and the C-terminus of the molecule. Unfortunately, no cysteine containing peptides were recovered from digests of peptide P1. However, since at 2 attempts at sequenator analysis of peptide CB2 no residue was identified at position 165 (whereas adjacent residues were easily detected) and knowing that PTH derivatives of cysteine are extremely difficult to detect, it is possible that residue 165 is cysteine. However, we appreciate that this point is not proven. As well as these sequence data, an elastase digest of total HMG 1 (unpublished) provided peptides ranging in size from 2–18 residues. These peptides were all sequences manually using the dansyl-Edman method and all have been placed in the sequence shown in fig.1. The elastase peptides accounted for a total of 180 amino acid residues and the cleavage positions for these peptides were all in agreement with the known specificity for elastase.

3.2. *The primary structure of HMG 2*

In both the peptic digest and tryptic digest of succinylated protein, peptides were obtained for HMG 2 which were analogous to those obtained for HMG 1. This is not surprising considering the strong sequence homology between HMG 1 and 2 (see below). For the peptic peptides P1 and P2 the approach was the same as used for HMG 1, namely, further digestion with trypsin, thermolysis and V8-protease, followed by peptide purification. Again, peptides were obtained which almost totally accounted for the sequence data that was obtained by sequenator analysis, but again the same 2 regions of indeterminate sequence remained. The only difference between the total data obtained for HMG 1 and 2 was that the cysteine containing peptide which was isolated from peptide P1 and HMG 1 was not isolated from the corresponding P1 peptide from HMG 2.

4. Discussion

Comparison of the sequence data for HMG 1 and 2 is shown in fig.2. Although for both proteins 2 short regions of sequence remain indeterminate, the overall architecture of both molecules is immediately apparent. Considering the highly homologous nature of these 2 proteins, it seems likely that the small difference in the total number of residues determined for both proteins (259 in HMG 1, 256 in HMG 2) is probably a consequence of the methods used to determine the compositions of the undetermined regions (subtracting a known sequence from an amino acid composition) and not an indication of different lengths of the 2 proteins. It has been known since the original characterisation studies on HMG 1 and 2 that the two proteins have very similar structures [24], and this is borne out by the sequence data. A comparison of the primary structures of HMG 1 and 2 is shown in fig.2. Of the 188 residues shown in sequence in fig.2, only 37 differences occur between the 2 proteins, and 19 of these changes can be considered as conservative. HMG 1 and 2 are therefore highly homologous proteins. Comparison with sequence data for other chromosomal proteins shows that no extensive regions of sequence homology exist between HMG 1 and 2 and the histones, or any of the other HMG proteins.

One of the interesting features of HMG 1 and 2 is the fact that >50% of their amino acids are charged, and the sequence data have revealed an asymmetric distribution of these charged groups. The most intriguing observation in both proteins is the presence of a continuous sequence of 35–40 aspartic and glutamic acid residues in the C-terminus of the molecule (residues 191–232 in HMG 1). This region is referred to as the HGA region (High Glutamic and Aspartic), and is preceded in both proteins by a group of basic amino acids. The HGA region in both proteins HMG 1 and HMG 2 must obviously play an important role in the function of these proteins, but the exact function of the HGA region has yet to be determined. In contrast to this highly acidic region, the basic amino acids are fairly evenly distributed, although small clusters of basic amino acids do occur in both proteins, e.g., the sequences:

Arg–Glu–Glu–His–Lys–Lys–Lys–His (residues 23–30 in HMG 1 and 2);

Lys–Lys–Gly–Lys–Lys–Lys (residues 92–97 in HMG 2); and

Lys–Ser–Lys–Lys–Lys–Lys (residues 186–192 in HMG 1).

However, no extended regions of basic residues, similar to the highly grouped acidic residues of the HGA peptide, exist in the molecule. Clusters of hy-

drophobic amino acids such as Ala–Phe–Phe–Leu–Phe–Ala (residues 108–113) and Tyr–Ala–Phe–Phe–Val–Gln (residues 15–20) also occur in the N-terminal half of both molecules. Such hydrophobic regions may well represent sites of protein–protein interactions much in the same way as the hydrophobic regions of the histones are thought to bind together in the histone octamer core of the nucleosome. The overall picture of the HMG 1 and 2 molecule is, therefore, one of a basic N-terminal half containing clusters of basic and hydrophobic amino acids but few acidic amino acids, whereas the C-terminal region contains the majority of the acidic residues in a continuous sequence. Side chain modification by acetylation, methylation and ADP-ribosylation has been reported for HMG 1, and sites of acetylation (at lysine residues 2 and 11) have been positively identified [25]. This reversible modification of lysine by acetylation in the N-terminal region is comparable with that observed with the histones.

In the light of the above sequence data it is difficult to imagine HMG 1 and 2 as having different functions. A reasonable explanation for the sequence similarity between HMG 1 and 2 would be that they have evolved from a common ancestral gene. Gene duplication followed by the separate evolution of both genes would give rise to two proteins with closely related sequences. Multiple forms of a chromosomal protein produced by microheterogeneity in the amino acid sequence has, of course, already been demonstrated in the case of histone H1 [22,23], although in this case as well the reasons for the heterogeneity are not known. Unfortunately, the absence of a function for the HMG proteins other than 'structural' precludes the comparison of these 2 proteins in some form of biological assay. Until such time as this can be done, the reason for the presence in chromatin of 2 such closely-related proteins will remain a mystery. However, if HMG 1 and 2 do have different functions, then the nature of this difference is most likely to relate to the structure preceding the HGA sequence (residues 174–190). This region shows the greatest density of differences between the two proteins (9 changes in 13 residues), and in particular involves 3 proline residues. Additionally, there is an extra proline residue at the beginning of the HGA sequence in HMG 2 and one further proline residue within the HGA sequence. These differences will produce considerable conformational changes between the 2 proteins in this region, and will almost certainly involve the HGA peptide in both proteins being held in different configurations with respect to the rest of the molecule. It is interesting to note that the 11 remaining proline residues found in the remainder of the molecule are all in identical positions in HMG 1 and 2, i.e., no gross conformational changes have been brought about in the rest of the molecule by the addition or deletion of proline residues. Otherwise the majority of all other differences occur as single changes and are fairly regularly distributed throughout the protein.

Acknowledgements

This work was supported by a grant to the Chester Beatty Research Institute (Institute of Cancer Research, Royal Cancer Hospital) from the Medical Research Council.

References

[1] McGhee, J. D. and Felsenfeld, G. (1980) Ann. Rev. Biochem 49, 1115–1156.
[2] Walker, J. M., Goodwin, G. H., Smith, B. J. and Johns, E. W. (1980) in: Comprehensive Biochemistry (Stotz, E. H. and Neuberger, A. eds) vol. 19B, Elsevier/North-Holland, Amsterdam, New York.
[3] Rabbani, A., Goodwin, G. H. and Johns, E. W. (1978) Biochem. Biophys. Res. Commun. 81, 351–358.
[4] Sterner, R., Boffa, L. C. and Vidali, G. (1978) J. Biol. Chem. 253, 3830–3836.
[5] Watson, D. C., Peters, E. H. and Dixon, G. H. (1977) Eur. J. Biochem. 74, 53–60.
[6] Rabbani, A., Goodwin, G. H., Walker, J. M., Brown, E. and Johns, E. W. (1980) 109, 294–298.
[7] Spiker, S., Mardian, J. K. W. and Isenberg, I. (1978) Biochem. Biophys. Res. Commun. 82, 129–135.
[8] Alfageme, C. R., Rudkin, G. T. and Cohen, L. H. (1976) Proc. Natl. Acad. Sci. USA 73, 2038–2042.
[9] Goodwin, G. H., Nicolas, R. H. and Johns, E. W. (1975) Biochim. Biophys. Acta 405, 280–291.
[10] Goodwin, G. H., Rabbani, A., Nicolas, R. H. and Johns, E. W. (1977) FEBS Lett. 80, 413–416.
[11] Sanders, C. (1975) PhD Thesis, London University.
[12] Goodwin, G. H., Woodhead, L. and Johns, E. W. (1977) FEBS Lett. 73, 85–88.
[13] Walker, J. M., Hastings, J. R. B. and Johns, E. W. (1977) Eur. J. Biochem. 76, 461–468.
[14] Walker, J. M., Goodwin, G. H. and Johns, E. W., (1979) FEBS Lett. 100, 394–398.
[15] Levy, W. B. and Dixon, G. H. (1978) Nucl. Acids. Res. 5, 4155–4163.

[16] Weisbrod, S. and Weintraub, H. (1979) Proc. Natl. Acad. Sci. USA 760, 630–634.
[17] Walker, J. M., Parker, B. P. and Johns, E. W. (1978) Int. J. Pept. Prot. Res. 12, 269–276.
[18] Walker, J. M., Gooderham, K. and Johns, E. W. (1979) Biochem. J. 181, 659–665.
[19] Hunkapiller, M. W. and Hood, L. E. (1978) Biochemistry 17, 2124–2129.
[20] Walker, J. M., Hastings, J. R. B. and Johns, E. W. (1978) Nature 271, 281.
[21] Walker, J. M., Goodwin, G. H., Johns, E. W., Wietzes, P. and Gaastra, W. (1977) Int. J. Pept. Prot. Res. 9, 220–223.
[22] Kinkade, J. M. and Cole, R. D. (1966a) J. Biol. Chem. 241, 5790–5797.
[23] Kinkade, J. M. and Cole, R. D. (1966b) J. Biol. Chem. 241, 5798–5805.
[24] Walker, J. M., Goodwin, G. H. and Johns, E. W. (1976) Eur. J. Biochem. 62, 461–469.
[25] Sterner, R., Vidali, G. and Allfrey, V. (1979) J. Biol. Chem. 254, 11577–11583.